Published in final edited form as:
J Am Chem Soc. 2007 October 10; 129(40): 12310–12319. doi:10.1021/ja0744899.

Quantitative Microarray Profiling of DNA-Binding Molecules

James W. Puckett[†], **Katy A. Muzikar**[†], **Josh Tietjen**[‡], **Christopher L. Warren**[‡], **Aseem Z. Ansari**[*,‡], and **Peter B. Dervan**[*,†]

[†]Contribution from the Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125

[‡]Department of Biochemistry and Genome Center, University of Wisconsin, Madison, Wisconsin 53706

Abstract

A high-throughput Cognate Site Identity (CSI) microarray platform interrogating all 524 800 10-base pair variable sites is correlated to quantitative DNase I footprinting data of DNA binding pyrrole-imidazole polyamides. An eight-ring hairpin polyamide programmed to target the 5 bp sequence 5′-TACGT-3′within the hypoxia response element (HRE) yielded a CSI microarray-derived sequence motif of 5′-WWACGT-3′ (W = A,T). A linear β-linked polyamide programmed to target a $(GAA)_3$ repeat yielded a CSI microarray-derived sequence motif of 5′-AARAARWWG-3′ (R = G,A). Quantitative DNase I footprinting of selected sequences from each microarray experiment enabled quantitative prediction of K_a values across the microarray intensity spectrum.

Introduction

Cell-permeable small molecules which bind specific DNA sequences and are able to interfere with protein–DNA interfaces would be useful in modulating eukaryotic gene expression. For targeting the regulatory elements of eukaryotic genes, knowledge of the preferred binding landscape of the ligand and the energetics of each site would guide gene regulation studies. Pyrrole–imidazole polyamides are a class of cell permeable oligomers which can be programmed, based on simple aromatic amino acid pairing rules, to bind a broad repertoire of DNA sequences.[1] Knowledge of polyamide match sites has allowed us to pursue the characterization of the equilibrium association constants and, hence, free energies of hairpin polyamides for cognate DNA sites by quantitative footprint titration methods. Despite the predictive power of simple pairing rules, the sequence dependent variability of DNA minor groove shape affords significant variability in the range of affinities for match as well as all formal single and double base pair mismatch sites.[1]

Quantitative Footprint Titrations

Characterization of polyamide binding preferences has been studied using quantitative DNase I footprinting titrations, affording binding isotherms that enable rigorous determination of the equilibrium association constant, K_a.[2] resolution of footprinting is

dervan@caltech.edu; ansari@biochem.wisc.edu.

Supporting Information Available: Experimental materials and methods, polyamide synthesis, plasmid preparation, and DNase I footprint titration details. Histograms of the frequencies of microarray intensities and fractional standard deviations. Additional K_a vs CSI microarray intensity plots and K_a-weighted sequence logos of polyamides **2** and **4**. This material is available free of charge via the Internet at http://pubs.acs.org/.

conservatively limited to association constants of 2-fold difference or greater. Polyamide binding preferences have frequently been interrogated using DNA fragments roughly 100 bp in size containing as many as four 6–10 bp binding sites, which are identical with the exception of a single position that iteratively exhibits A•T, T•A, C•G, and G•C base pairs. Each binding site is interspersed with an 8 or more base pair spacer region to prevent interaction between the binding sites.[3] Obtaining high quality data limits a ^{32}P end-labeled DNA fragment to four unique binding sites due to the resolving power of a polyacrylamide gel in a quantitative footprint titration. While DNase I footprinting has enabled the elucidation of a binding code for hairpin polyamides, a relatively limited set of binding sites has been studied. To comprehensively interrogate all four encoded positions of an eight-ring hairpin polyamide, one would need 136 unique binding sites. In addition, interrogation of the base pairs flanking the polyamide core would necessitate 2080 (for 6 bp total) or 32 896 (for 8 bp total) binding sites.

CSI Microarray Platform

Several high-throughput platforms have been developed to characterize the binding properties of ligand–DNA interactions.[4] Of these, two have been used to explicitly study the binding preferences of polyamides. The fluorescence intercalator displacement assay has interrogated polyamide binding to 512 unique 5 bp sequences in a microplate format.[4b] The more recently developed cognate site identifier (CSI) microarray platform presents all 32 896 unique eight-mers (scalable to all unique ten-mers) to fluorescently labeled polyamides, enabling an unbiased interrogation of binding preference.[4c] By coupling DNase I footprinting with the CSI microarray data, the binding affinities (K_a values) of DNA-binding molecules for a significantly larger number of DNA sequences could be determined (Figure 1). To date, CSI microarray intensities of hairpin polyamide–Cy3 conjugates have been linearly correlated to the K_a values of unlabeled polyamides.[4c] We will examine whether this relationship between DNase I footprint titration-derived K_a values for Cy3-labeled polyamides and the corresponding microarray data remains true for additional polyamide binding architectures. Because the Cy3–polyamide conjugate may alter sequence specificity when compared with its biologically active counterpart, the sequence specificities of fluorophore-labeled polyamide and the biologically relevant polyamide will also be determined.

A CSI microarray harbors immense sequence specificity data; determining how to best represent this data is critical. The first reported CSI work[4c] represented binding preferences as a sequence logo[5] derived from several motif-finding algorithms[6] that searched the highest *Z*-score bins (the ~300 highest intensities on the array), assigning equal weight to each sequence. It also examined the relative abundance of each sequence motif mutation within its respective *Z*-score bin.[4c] In this paper we observe that K_a-weighting sequence motifs does not alter the sequence logo appreciably. In addition, a comprehensive single base pair mutational analysis is performed, which quantifies the specificities encoded by the polyamide at each position the polyamide interacts with DNA.

Two Cy3-labeled polyamides of biological interest[7] are examined on a CSI microarray that displays all unique 10 base pair DNA sequences. These polyamides include a hairpin structure whose sequence specificities can be predicted from the extensive DNase I footprinting data characterizing other pyrrole–imidazole polyamides[1] and a linear β-linked structure whose sequence specificity is less well understood.[8] In order to correlate the CSI relative affinities (intensities) to absolute affinities (K_a values), DNase I footprinting was performed on a subset of these sequences for both the Cy3–polyamide conjugates and the related, unlabeled polyamides of known biological activity.

Results and Discussion

Polyamide Design

Two polyamide core sequences have been chosen as representative of both hairpin and linear β-linked polyamide architectures. These core recognition sequences exhibit biologically significant roles, modulating transcription in cell culture experiments.[7] Hairpin polyamides **1** and **2** (Figure 2) were selected based on results from a project in which a polyamide–fluorescein conjugate, Ct-Py-Py-Im-(*R*)-^{H2N}-γ-Py-Im-Py-Py-Dp-FITC (**1**) displaced hypoxia inducible factor-1α (HIF-1α) from the hypoxia response element (HRE) of the vascular endothelial growth factor (VEGF) gene, downregulating VEGF expression 60% in cell culture experiments.[7a,b] This eight-ring hairpin was programmed to bind the sequence 5′-WTWCGW-3′ (W = A,T).[1,3c] In particular, polyamide **1** was shown to bind the HRE sequence, 5′-TACGTG-3′, on the VEGF promoter by footprint titration.[7a,b] The Cy3 moiety was conjugated (**2**) at the same position as fluorescein for **1** to best mimic the binding properties between the two polyamides.

As with polyamide **1**, polyamide **3** (Figure 2) is known to bind its biologically relevant target. Polyamide **3**, Im-Py-β-Im-Py-β-Im-β-Dp, targets an intronic 5′-$(GAA)_n$-3′ repeat hyper-expansion, enabling 2.5-fold upregulation of the *frataxin* gene, whose deficiency causes the neurodegenerative disorder Friedreich's Ataxia.[7c] Limited knowledge about the linear β-linked class of polyamides[8] precludes the existence of binding rules. The linear β-linked architecture has the added complexity of binding in 1:1 and 2:1 ligand/DNA stoichiometries, and we would anticipate that this class will be generally less useful due to sequence promiscuity resulting from multiple binding modes. Its 1:1 binding preferences for purine tracts, such as $(GAA)_n$, likely reflect shape selectivity for sequences with narrow DNA minor groove conformations.[8c] In a 2:1 binding stoichiometry, polyamide **3** would be predicted to target 5′-WGCWGCWGCW-3′.[8a] Remarkably, relatively few genes are affected from cell culture studies of **3** suggesting that this polyamide may be specific for 5′-AAGAAGAAG-3′.[7c] The Cy3 fluorophore has been conjugated to the C-terminal 3,3′-diamino-*N*-methyldipropylamine tail (polyamide **4**).

CSI Microarray Design and Results

CSI microarrays were synthesized using maskless array synthesis (MAS) technology[9] to display all 524 800 unique 10-base pair sites in quadruplicate across six microarrays. Replicates of individual hairpins occur on separate microarrays. Each hairpin on the chip consists of a self-complementary palindromic sequence interrupted by a central 5′-GGA-3′ sequence to facilitate hairpin formation: 5′-GCGC-$N^1N^2N^3N^4N^5N^6N^7N^8N^9N^{10}$-GCGC-GGA-GCGC$N^{10'}N^{9'}N^{8'}N^{7'}N^{6'}N^{5'}N^{4'}N^{3'}N^{2'}N^{1'}$-GCGC-3′ (N = A,T,C,G). Previous experiments have found that 95% of the oligonucleotides on the array form duplexes.[4c]

Polyamides **2** and **4** were slowly titrated onto the arrays and imaged at each concentration until saturation of the highest intensity binding sites was observed, 10 nM and 175 nM concentrations, respectively, for **2** and **4**. After each small addition of polyamide, the arrays were washed prior to imaging. The data for each of the arrays were then normalized as previously described[4c] to give averaged sequence intensities of the 524 800 10-base pair sites for **2** and **4**. As found with previously reported CSI arrays,[4c] histograms of the probe intensities for **2** and **4** display a strong right-handed tail (Supporting Information Figure 1). The fractional standard deviations among probe replicates (standard deviation of replicates/average normalized intensity) average 0.15 ± 0.09 (polyamides **2** and **4**), for intensities exceeding 1×10^3 (Supporting Information Figure 2).

Plasmid Design

Three plasmids have been designed based on output from the CSI microarray intensities (Figure 3). Because of our interest in testing the dynamic range of the CSI assay in terms of the representative K_a values measured by a broad range of intensities, plasmids pKAM3 and pJWP17 were constructed to harbor binding sites of equal intensity spacing across a broad portion of each array's intensities, between highest and lowest intensities. The K_a values found using pKAM3 were clustered across the three highest intensities, necessitating further interrogation. Plasmid pKAM4 was designed to probe three additional intensities. A single binding site (**IIIa** and **Ib**) was held constant between pKAM3 and pKAM4 to enable interplasmid comparison of binding affinities. Because pJWP17 afforded K_a values broadly spaced across the intensity spectrum, no further study was pursued.

Since our goal is to directly compare footprinting-derived K_a values with CSI-array derived intensities, each plasmid binding site mimics the full 10 base pair binding site from the array in addition to two flanking base pairs on either side of the binding site: 5′-GC-$(N)_{10}$-GC-3′ (N = A,T,C,G). Attempts to fully replicate the 5′-GCGC-$(N)_{10}$-GCGC-3′ binding site from the array exhibited secondary structure formation when the respective amplicons were sequenced and separated by denaturing gel electrophoresis.

Quantitative DNase I Footprint Titrations: Affinity and Specificity Determination

Hairpin polyamides **1** and **2** were incubated each for 14 h with pKAM3 or pKAM4 prior to DNase I cleavage. These two polyamides were found to bind each of seven unique 10-base pair binding sites in the same rank order, preferentially binding 5′-TTTTACGTAA-3′ with affinities of 7.5×10^9 M^{-1} (**1**) and 4.5×10^9 M^{-1} (**2**) (Figure 4 and Table 1).

Replacing the flourescein dye on polyamide **1** with Cy3 (polyamide **2**) introduced an energetic penalty that ranged from 1.5- to 10-fold, with the minimum penalty occurring at the two highest CSI intensity binding sites (Table 1). Polyamide **2** differentiated the highest and lowest affinity binding sites by 70-fold, slightly more than the 50-fold differentiation found for the fluorescein-labeled polyamide **1**.

Linear β-linked polyamides **3** and **4** were each incubated for 14 h with pJWP17 prior to DNase I cleavage. They bound four unique 10-base pair sites in the same rank order, preferentially binding 5′-AAGAAGAAGT-3′ (Table 2 and Figure 5).

Appending the Cy3 dye to polyamide **3** had either no effect on affinity or reduced binding affinity as much as 30-fold (Table 2). Polyamide **3** bound all four binding sites over a 2400-fold range in affinity, eight times broader than that for polyamide **4**.

Calibrating Microarrays for K_a Prediction

Because DNase I footprinting enables the calculation of K_a and the direct comparison of four binding sites in a single assay, determining energetics data from CSI microarrays is crucial for understanding the global binding specificity of a polyamide. An eight-ring hairpin polyamide targeting 5′-WGWWCW-3′ (W = A,T) and characterized by quantitative DNase I footprinting, Im-Py-Py-Py-γ-Im-Py-Py-Py-β-Dp,[3c,10] has been compared to its Cy3-labeled counterpart studied on the CSI-array platform, demonstrating a linear relationship between intensity and K_a.[4c]

Because microarray intensity at a specific microarray feature should be proportional to the fractional occupancy of DNA at that feature, the relationship between equilibrium association constant (K_a) or dissociation constant (K_d) and background-normalized microarray intensity should be[11]

$$\text{Intensity}=c \times \Theta=c \times \frac{K_a[\text{PA}]}{1+K_a[\text{PA}]}=c \times \frac{[\text{PA}]}{K_d+[\text{PA}]} \quad (1)$$

In this relationship, Θ represents the fractional occupancy of DNA at a specific feature, *c*, a scalar to reflect that microarray intensity can vary with incident laser intensity, and [PA], the free polyamide concentration on the CSI array. The terms *c* and [PA] are solved for a curve fit to eq 1 using K_a values derived from DNase I footprint titrations and CSI microarray intensity data. Examining the limiting case where $[\text{PA}] \ll K_d$ one observes a simplification to eq 1:

$$\text{Intensity}=c \times \frac{[\text{PA}]}{K_d}=c \times [\text{PA}] \times K_a \quad (2)$$

Equation 2 represents the linear subset of the more general CSI intensity – K_a relationship described in eq 1. Fitting the footprinting data of polyamide **2** to its corresponding microarray intensities (Table 1) using eq 2 fits well (R^2 = 0.94). The linearized eq 2 does not, however, map intensity and K_a with high correlation for polyamide **4**. Fitting the data to eq 1 affords a significantly better fit (R^2 = 0.99), indicating that [PA] is not insignificant relative to the K_d of the highest intensity microarray data (Figure 6).[12,13]

The K_a-calibrated microarrays can subsequently be used to interpolate K_a values from normalized sequence intensities. K_a values are derived by rearranging eq 1 to present K_a as a function of microarray intensity:

$$K_a=\frac{\text{Intensity}}{[\text{PA}] \times (c-\text{Intensity})} \quad (3)$$

In the case where $[\text{PA}] \ll K_d$, eq 2 is rearranged to

$$K_a=\frac{\text{Intensity}}{[\text{PA}] \times c} \quad (4)$$

Correlating Binding Between Cy3-Labeled and Biologically Relevant Polyamides

While establishing a general K_a–intensity relationship for Cy3-labeled polyamides is a crucial first step toward global sequence interrogation of a core polyamide motif, it is equally important that the biologically relevant polyamide has sequence preferences that correlate with its Cy3-labeled counterpart. Scatter plots of polyamide **1** vs **2** and polyamide **3** vs **4** are best fit by a power relationship of $y = ax^n$, where (*x*,*y*) denotes the K_a values for (**1**, **2**) or (**3**, **4**) (Figure 7).[14] The R^2 between **1** and **2** is 0.87, and that between **3** and **4** is 0.78.

Sequence Analysis

To graphically represent the binding preferences of polyamides **2** and **4**, sequence logos have been generated (Figures 8 and 9).

In all cases, the motif finding program MEME[6a] was utilized to extract sequence motifs from the CSI binding intensities. The position specific probability matrices output by MEME were used as inputs to enoLOGOS[15] to generate a sequence logo.[16] The logo for

polyamide **2** was created by searching the ~2500 highest sequence intensities of the CSI microarray.[17] These data points span approximately a 3-fold range in K_a. The logo for polyamide **4** interrogated the 48 highest intensity sequences (a 7-fold range in K_a) of the CSI microarray.[18] We examined K_a-weighted sequence logos for both polyamides **2** and **4** and found minimal differences in the resulting logos (Supporting Information Figure 3).

The motif for polyamide **2** has the most information at a site width of six – 5′-WWACGT-3′ (Figure 8; W = A,T). The chlorothiophene/pyrrole pair (Ct/Py) specificity cannot be globally elucidated using polyamide **2** because of the palindromic nature of the ACGT binding site core. It is evident that the core does specify 5′-ACG-3′ using Py/Py, Py/Im, and Im/Py pairings, respectively. Polyamide **3** specifies 9 base pairs based on MPE footprinting data (unpublished). Polyamide **4** elicits a 9 bp motif that is best represented as 5′-AARAARWWG-3′ (Figure 9; R = G,A and W = A,T). Previous work would suggest that Im may have no sequence preferences within linear β-linked polyamides,[8] although this selection of 9 bp high affinity binding sites for **4** suggests at least G•C or A•T specificity, consistent with microarray data from Friedreich's Ataxia cell culture work.[7c]

Quantitative Profiling of Single Base Pair Mismatches

While sequence logos provide a visual representation of sequence specificity, traditional studies on polyamides quantitate the specificity of a ring pairing at a selected base pair. We have examined a comprehensive single base pair mutational analysis of both polyamides **2** and **4** using K_a values interpolated from the calibrated CSI microarrays (Tables 3 and 4).[19]

Because the motif finding algorithm MEME found 5′-WWACGT-3′ (W = A,T) as a preferred binding sequence for polyamide **2**, we utilized this core sequence for mutational studies. Additionally, because of the 5′-ACGT-3′ palindromic element of this binding site, we have isolated only binding sites containing 5′-WWWWWWACGT-3′ and their mutant counterparts to preclude analyzing variants where the polyamide may be rotated 180° from the presumed orientation. To determine a K_a for 5′-WWWWWWACGT-3′ (for example), the geometric mean of all microarray binding sites containing this motif was found. Walking from 5′ to 3′ on 5′-$W^1W^2A^3C^4G^5T^6$-3′, we observe that there is 3-fold specificity for W versus S (S = C,G) at position 1 (occupied by the linker). At position 2 (Ct/Py pair), there is 20-fold specificity for W versus S but minimal for T•A versus A•T. The previous study of Ct/Py specificity noted only modest specificity for T•A versus A•T.[3c] Position 3 (a Py/Py pair) confirms the previously observed W over S specificity.[1] At position 4 (a Py/Im pair) the polyamide encodes the greatest specificity with preference for C•G versus A•T, T•A, or G•C. It is likely that this preference is at least 20-fold. At position 5, polyamide **2** appears to exhibit less specificity than would be predicted for an Im/Py ring pair, binding almost as well to A•T as to G•C.[1] The polyamide "turn unit," position 6, confirms a strong preference for W over S.[1] Through this quantitative study, we observe four strongly encoded binding positions, italicized in 5′-W*WWC*G*W*-3′. The discrepancy between the observed sequence logo, as found by MEME, and the suggested specificity by a single base pair mutation study likely stems from (i) the examination of all sequences in the single base pair mutation as compared to only a subset for the sequence logo, (ii) the assumption by the logo of independence of base pair–polyamide interaction at each position, and (iii) the examination in the single base pair mutation of the average K_a of a group of sequences containing a specified motif.

In conjunction with the sequence logo for polyamide **2**, the CSI array analysis validates the sequence specificity programmed by the aromatic amino acid ring pairs. The extensive DNase I footprinting data on eight-ring and six-ring hairpin polyamides, while limited on the scale of a CSI microarray, enabled the creation of pairing rules that are remarkably general.[1] It is evident from the microarray that Im/Py and Py/Im ring pairs offer the greatest

specificity for a single base pair, while Py/Py, Ct/Py, and the "turn unit" afford general W specificity. While the Ct/Py ring pair conferred minimal specificity for T•A versus A•T, its W specificity is likely an improvement over the use of a Py/Py ring pair, which at the N-terminus of an eight-ring hairpin polyamide exhibits specificity for A•T, T•A, and G•C versus C•G.[3c] The sequence specificity of **2** correlates remarkably well with the 5′-ACGT-3′ specificity of echinomycin, [20] also known to affect VEGF expression in cell culture.[21]

The examination of polyamide **4** marks the most comprehensive sequence specificity study of a linear β-linked polyamide since the original examination of the binding specificity for Im-β-Im-Py-β-Im-β-Im-Py-β-Dp.[8a,b] In the 5′-$A^1A^2R^3A^4A^5R^6W^7W^8G^9$-3′ sequence (R = G,A; W = A,T), positions 4, 5, and 7, each containing either a Py or a β, exhibit the greatest specificity for W over S (S = C,G). Intriguingly, the β at position 4 prefers A•T over T•A, an unexpected specificity. The sequence logo for polyamide **4** indicates that Im has a modest preference for G•C or A•T over other base pairings; in this mutational study, however, imidazole is generally degenerate. The wide range of K_a values comprising each motif (high geometric standard deviation) make the statistical significance of any specificities under 4 relatively small. In general, the geometric standard deviations for polyamide **4** were higher than those for polyamide **2**, when including only those table entries for polyamide **2** in which each K_a value was composed of all instances of the motif. One potential source of the increased standard deviation in binding affinities is the single variable base flanking the nine base pair binding site for polyamide **4**. Because the minor groove width is a potentially important contributor to binding affinity and specificity for the linear β-linked class of polyamides,[8c] a single variable, flanking base is unlikely to enable comprehensive interrogation of the global set of sequence-dependent DNA microstructures. As with polyamide **2**, the discrepancies observed between the sequence logo of polyamide **4** and the comprehensive single base pair mutational analysis likely stem from similar causes.

With the sequence logo (approximated as 5′-AARAARWWG-3′) as a snapshot of the highest affinity binding sites for polyamide **4** ($K_a \approx 5 \times 10^8$ to 3.3×10^9 M^{-1}) and the footprint titration binding isotherms for determining DNA binding mode, we confirm a preference for the 1:1 binding stoichiometry. Previous data characterizing the linear β-linked polyamide Im-β-Im-Py-β-Im-β-Im-Py-β-Dp demonstrated a 30-fold energetic preference for the 1:1 versus 2:1 binding stoichiometry, presumably due to the increased entropic cost of the 2:1 binding mode.[8a] It is remarkable that polyamide **3** exhibited specificity for upregulation of the *frataxin* gene in cell culture,[7c] since the sequence preference for **4** was not overwhelmingly 5′-AAGAAGAAG-3′. Two possible explanations for this observation are (i) that multiple binding events in the genome have marginal effects on transcription and that the specificity is amplified by the GAA repeat expansion in Friedreich's Ataxia or (ii) that many of the sequences described by 5′-AARAARWWG-3′ exist in higher order chromosomal structures that cannot be targeted by polyamide **3**.

Suggestions for Microarray Usage

In the case where the free ligand concentration is small relative to the K_d for each binding site on the CSI microarray, a linear K_a–intensity relationship is observed. The binding profiles examined for polyamide **2** and for previously studied molecules are examples of linear K_a–intensity relationships.[4c] For the highest intensity sites also studied by DNase I footprinting (Figure 6a), the CSI microarray experiment contains greater resolving power and can differentiate K_a values that are indistinguishable by quantitative DNase I footprint titrations. In this example, as CSI intensity data approach ε, small changes in intensity yield large changes in predicted K_a. Because the characterization of DNA-binding ligands is most concerned with defining a perfect match site, this limitation is minor. CSI data for polyamide **2** conservatively enables distinguishing a 50-fold range of K_a values, thus encompassing the majority of single base pair mismatch specificities.

In the case where the free ligand concentration is comparable to the K_d, a nonlinear K_a–intensity relationship is observed. The binding profile for polyamide **4** marks an example of a CSI microarray studied compound that occurs outside the linear range of eq 1. In this case, clustered high-intensity data points can span a broad range of K_a values (Figure 6b). The error inherent to the CSI microarray analysis is thus amplified when K_a values in this high CSI intensity region are interpolated.

Because of the gradual polyamide titration onto the array, it should be possible to capture snapshots of both polyamide saturation within the linear K_a–intensity region for the highest affinity binding sites on the microarray and binding site saturation enabling lower intensity data points to fall within the higher precision linear K_a–intensity region. Such titration may enable high precision K_a data to be extracted from all intensities of the microarray.

The sequence logos presented in this paper represent a snapshot of a binding profile for the highest affinity binding sites by a dye-labeled ligand. The polyamide core dictates the majority of the binding specificity revealed by CSI microarray analysis – the presence of a Cy3 label may reduce affinity to a binding site relative to its unlabeled counterpart but does not alter the rank order of binding preferences. Complementing the graphical image of a sequence logo, the comprehensive single base pair mutational analysis afforded by the extensive microarray data quantitates one's understanding of the polyamide sequence preferences.

Conclusion

Correlating the sequence preference landscape present on the CSI microarray to quantitative footprinting enables energetic studies using global binding information. This capacity marks a significant forward step for the field of small molecule•DNA recognition and enables the comprehensive interrogation of DNA binding small molecules to be better understood. The elucidation of 5′-WWACGT-3′ as the binding site for **2** confirmed the previously established pairing code for hairpin polyamides, and the determination of 5′-AARAARWWG-3′ for **4** helps explain the specificity it exhibited in cell culture. The correlation between a Cy3-labeled polyamide and an unlabeled polyamide of biological interest means that these motifs well approximate the binding profiles for **1** and **3**, respectively. DNase I footprinting-calibrated CSI microarrays have been shown to be an effective technique for determining the binding affinities of DNA-binding ligands for a vastly expanded repertoire of DNA sequences, and we envision them to be a critical tool for reliably determining sequence specificity for other ligands in the future.

Experimental Section

Materials, methods, synthesis, plasmid preparation, and DNase I footprinting procedures are found in the Supporting Information.

Microarray Procedures

Microarrays were synthesized by using a Maskless Array Synthesizer (NimbleGen Systems, Madison, WI). Homopolymer (T_{10}) linkers were covalently attached to monohydroxysilane glass slides. Oligonucleotides were then synthesized on the homopolymers to create a high-density oligonucleotide microarray. The array surface was derivatized such that the density of oligonucleotides was sufficiently low within the same feature so that no one oligonucleotide would hybridize with its neighbors. Four copies of each hairpin containing a unique 10 bp site (5′-GCGC-$N^1N^2N^3N^4N^5N^6N^7N^8N^9N^{10}$-GCGC-GGA-GCGC-$N^{10'}N^{9'}N^{8'}N^{7'}N^{6'}N^{5'}N^{4'}N^{3'}N^{2'}N^{1'}$-GCGC-3′) required a total of 2 099 200 features, divided among six microarrays.

Binding Assay

Microarray slides were immersed in 1x PBS and placed in a 90 °C water bath for 30 min to induce hairpin formation of the oligonucleotides. Slides were then transferred to a tube of nonstringent wash buffer (saline/sodium phosphate/EDTA buffer, pH 7.5/0.01% Tween 20) and scanned to check for low background (<200 intensity). Microarrays were scanned by using an Axon 4000B, and the image files were extracted with GENEPIX PRO Version 3.0 (Axon Instruments, Foster City, CA).

Polyamide Binding

Microarrays prepared as above were placed in the microarray hybridization chamber and washed twice with nonstringent wash buffer. Polyamide was diluted to 10 nM (for **2**) or 175nM (for **4**) in Hyb buffer (100 mM Mes/1 M NaCl/20 mM EDTA, pH 7.5/0.01% Tween 20). Polyamide was then added to the hybridization chamber and incubated at room temperature for 1 h. Finally, the microarrays were washed twice with nonstringent wash buffer and scanned.

Data Processing

For each replicate, global mean normalization was used to ensure the mean intensity of each microarray was the same. Local mean normalization[22] was then used to ensure that the intensity was evenly distributed throughout each sector of the microarray surface. Outliers between replicate features were detected by using the *Q* test at 90% confidence and filtered out. The replicates were then quantile-normalized[23] to account for any possible nonlinearity between arrays. Duplicate features were then averaged together. The median of the averaged features was subtracted to account for background.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health (GM27681 to P.B.D. and A.Z.A). We thank the Beckman Institute Sequence Analysis Facility for DNA sequencing.

References

1. Im/Py targets G•C; Py/Py targets A•T and T•A; and Ct/Py targets T•A. (a) Dervan PB, Edelson BS. Curr. Opin. Struct. Biol. 2003; 13:284–299. [PubMed: 12831879] (b) Hsu CF, Phillips JW, Trauger JW, Farkas ME, Belitsky JM, Heckel A, Olenyuk BZ, Puckett JW, Wang CCC, Dervan PB. Tetrahedron. 2007; 63:6146–6151. [PubMed: 18596841]
2. Trauger JW, Dervan PB. Methods Enzymol. 2001; 340:450–466. [PubMed: 11494863]
3. (a) Doss RM, Marques MA, Foister S, Chenoweth DM, Dervan PB. J. Am. Chem. Soc. 2006; 128:9074–9079. [PubMed: 16834381] (b) Marques MA, Doss RM, Foister S, Dervan PB. J. Am. Chem. Soc. 2004; 126:10339–10349. [PubMed: 15315448] (c) Foister S, Marques MA, Doss RM, Dervan PB. Bioorg. Med. Chem. 2003; 11:4333–4340. [PubMed: 13129569]
4. (a) For a review on Protein Binding Microarrays (PBMs), see: Bulyk ML. Methods Enzymol. 2006; 410:279–299. [PubMed: 16938556] (b) For a review on fluorescence intercalator displacement (FID) assays, see: Tse WC, Boger DL. Acc. Chem. Res. 2004; 37:61–69. [PubMed: 14730995] (c) For the initial report of cognate site identifier (CSI) microarrays, see: Warren CL, Kratochvil NCS, Hauschild KE, Foister S, Brezinski ML, Dervan PB, Phillips GN, Ansari AZ. Proc. Natl. Acad. Sci. U.S.A. 2006; 103:867–872. [PubMed: 16418267]
5. Schneider TD, Stephens RM. Nucleic Acids Res. 1990; 18:6097–6100. [PubMed: 2172928]

6. (a) Bailey, TL.; Elkan, C. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. Altman, R.; Brutlag, D.; Karp, P.; Lathrop, R.; Searls, D., editors. Menlo Park: AAAI Press; 1994. p. 28-36.(b) Liu XS, Brutlag DL, Liu JS. Nat. Biotechnol. 2002; 20:835–839. [PubMed: 12101404] (c) Hughes JD, Estep PW, Tavazoie S, Church GM. J. Mol. Biol. 2000; 296:1205–1214. [PubMed: 10698627]
7. (a) Olenyuk BZ, Zhang GJ, Klco JM, Nickols NG, Kaelin WG, Dervan PB. Proc. Natl. Acad. Sci. U.S.A. 2004; 101:16768–16773. [PubMed: 15556999] (b) Nickols NG, Jacobs CS, Farkas ME, Dervan PB. Nucleic Acids Res. 2007; 35:363–370. [PubMed: 17175539] (c) Burnett R, Melander C, Puckett JW, Son LS, Wells RD, Dervan PB, Gottesfeld JM. Proc. Natl. Acad. Sci. U.S.A. 2006; 103:11497–11502. [PubMed: 16857735]
8. (a) Dervan, PB.; Urbach, AR. Essays in Contemporary Chemistry. Quinkert, G.; Kisakürek, MV., editors. Zurich: Verlag Helvetica Chimica Acta; 2000. p. 327-339.(b) Urbach AR, Dervan PB. Proc. Natl. Acad. Sci. U.S.A. 2001; 98:4343–4348. [PubMed: 11296283] (c) Urbach AR, Love JJ, Ross SA, Dervan PB. J. Mol. Biol. 2002; 320:55–71. [PubMed: 12079334] (d) Marques MA, Doss RM, Urbach AR, Dervan PB. HelV. Chim. Acta. 2002; 85:4485–4517.
9. Singh-Gasson S, Green RD, Yue YJ, Nelson C, Blattner F, Sussman MR, Cerrina F. Nat. Biotechnol. 1999; 17:974–978. [PubMed: 10504697]
10. (a) Trauger JW, Baird EE, Dervan PB. Nature. 1996; 382:559–561. [PubMed: 8700233] (b) Trauger, JW. Ph.D. Thesis. California Institute of Technology; 1999.
11. For derivation of these equations, see the Supporting Information for Bulyk ML, Huang XH, Choo Y, Church GM. Proc. Natl. Acad. Sci. U.S.A. 2001; 98:7158–7163. [PubMed: 11404456]
12. (a) $c \times [PA]$ was 1.7×10^{-5} for polyamide **2**. (b) c was 80.2×10^{3} and [PA] was 5.5×10^{-9} M for polyamide **4**. (c) To view plots reflecting the same curve fits of Figure 6 on a log–log scale, please see Supporting Information Figure 3.
13. Although the data for polyamide **2** (Figure 6a) maps intensity and K_a values using the linearized equation 2, this fit is distinct from that obtained by fitting the data to a line of the form $y = mx + b$, which includes an intensity-axis intercept term. While very small in this case, the differences in the slopes and intercepts of the lines may indicate error both in the background correction of the microarray and in the DNase I footprinting data. To correct for this possibility, we propose the use of an error term, ε, that would modify eqs 1 and 2 to the following: Intensity $= c \times \Theta + \varepsilon = c \times \{K_a[PA]\}/\{1 + K_a[PA]\} + \varepsilon = c \times \{[PA]\}/\{K_d + [PA]\} + \varepsilon$ (eq 1e) and Intensity $= c \times [PA]/K_d + \varepsilon = c \times [PA] \times K_a + \varepsilon$ (eq 2e). When fitting the intensity and K_a data for polyamide **2** to the modified equation 2e, one finds a marginally improved fit ($R^2 = 0.97$), although the curve fit for polyamide **4** using equation 1e is unimproved ($R^2 = 0.99$). For polyamide **2**, $c \times [PA] = 1.5 \times 10^{-5}$ and $\varepsilon = 5.5 \times 10^{3}$. For polyamide **4**, $c = 81.2 \times 10^{3}$, [PA] $= 5.7 \times 10^{-9}$ M, and $\varepsilon = -1.1 \times 10^{3}$.
14. For the relationship between polyamides **1** and **2**, $a = 0.0253$ and $n = 1.115$. For the relationship between polyamides **3** and **4**, $a = 349.83$ and $n = 0.637$.
15. Workman CT, Yin YT, Corcoran DL, Ideker T, Stormo GD, Benos PV. Nucleic Acids Res. 2005; 33:W389–W392. [PubMed: 15980495]
16. Figure 8 utilized 10 variable bases and contained a background GC content of 50%; Figure 9 utilized 10 variable bases and two fixed bases, each flanking the 5′ and 3′ portion of the variable region, and contained a background GC content of 58%. These background GC content corrections were utilized in the motif searching parameters.
17. There are 1258 occurrences of a full 6 bp match sequence, TTACGT. Double this number of highest intensity sequences was also searched yielding only modest changes in the data. The sequence logo is reported for the 2516 highest intensity sequences.
18. There are 24 occurrences of a full 9 bp match sequence, AAGAAGAAG on the microarray. Double this number of highest intensity sequences was searched in addition to searching only the 24 highest intensities, yielding only small changes in the data. The sequence logo reported contains the 48 highest intensity sequences.
19. To convert intensity to K_a, we have included the error term ε in our calculations. This gives modified versions of eqs 3 and 4, $K_a = (\text{Intensity} - \varepsilon)/\{[PA] \times (c - \text{Intensity} + \varepsilon)\}$ and $K_a = (\text{Intensity} - \varepsilon)/\{[PA] \times c\}$, respectively.
20. Van Dyke MM, Dervan PB. Science. 1984; 225:1122–1127. [PubMed: 6089341]

21. Kong D, Park EJ, Stephen AG, Calvani M, Cardellina JH, Monks A, Fisher RJ, Shoemaker RH, Melillo G. Cancer Res. 2005; 65:9047–9055. [PubMed: 16204079]
22. Colantuoni C, Henry G, Zeger S, Pevsner J. Bioinformatics. 2002; 18:1540–1541. [PubMed: 12424128]
23. Bolstad BM, Irizarry RA, Astrand M, Speed TP. Bioinformatics. 2003; 19:185–193. [PubMed: 12538238]

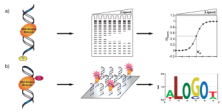


Figure 1.
(a) Quantitative DNase I footprinting gives rise to a defined equilibrium association constant at a specified binding site for a given DNA binding molecule. (b) The CSI microarray platform gives rise to relative binding preferences of an entire sequence space for the same molecule with a sequence logo as a standard summary output.

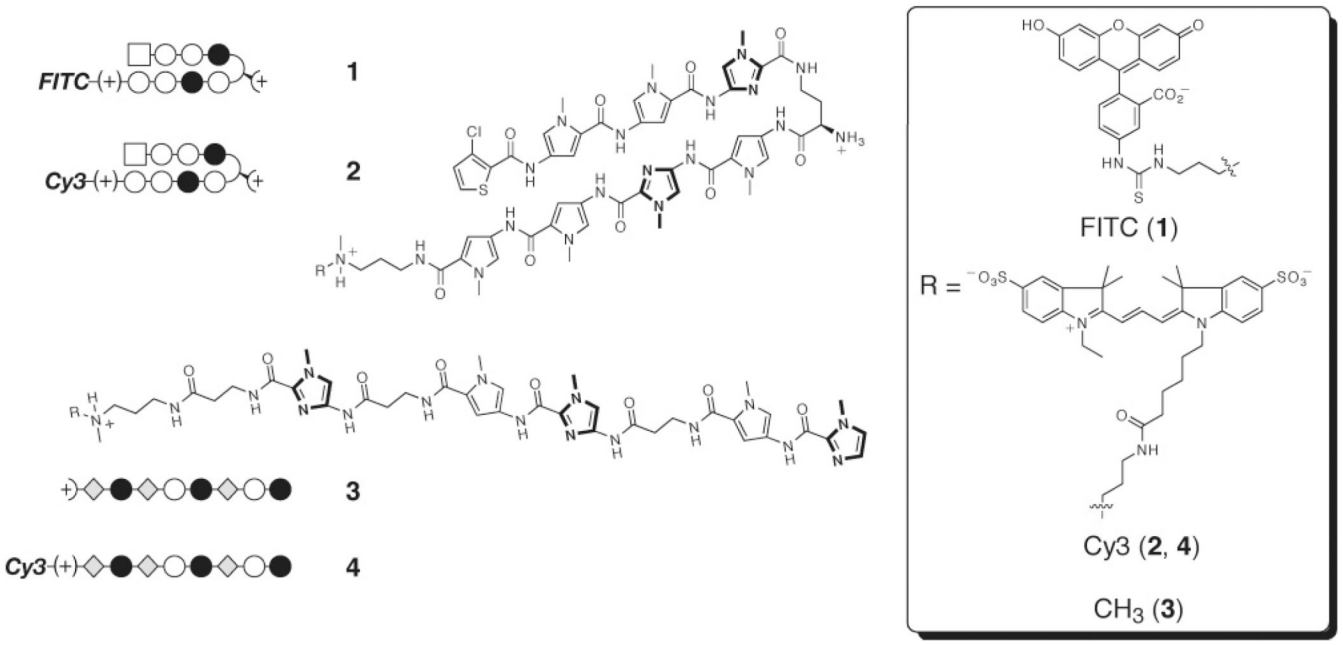


Figure 2.
Hairpin polyamides **1** and **2** targeted to the hypoxia response element (HRE), 5′-TACGTG-3′. Linear β-linked polyamides **3** and **4** targeted to GAA repeats in Friedreich's Ataxia.

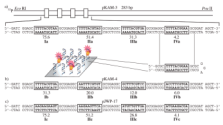


Figure 3.
Insert sequences utilized for plasmids, with binding sites boxed, labeled with their corresponding CSI array intensity, and numbered. (a) pKAM3 is shown, in addition to a microarray schematic demonstrating the relationship between the plasmid and a selected microarray sequence. (b) pKAM4. (c) pJWP17.

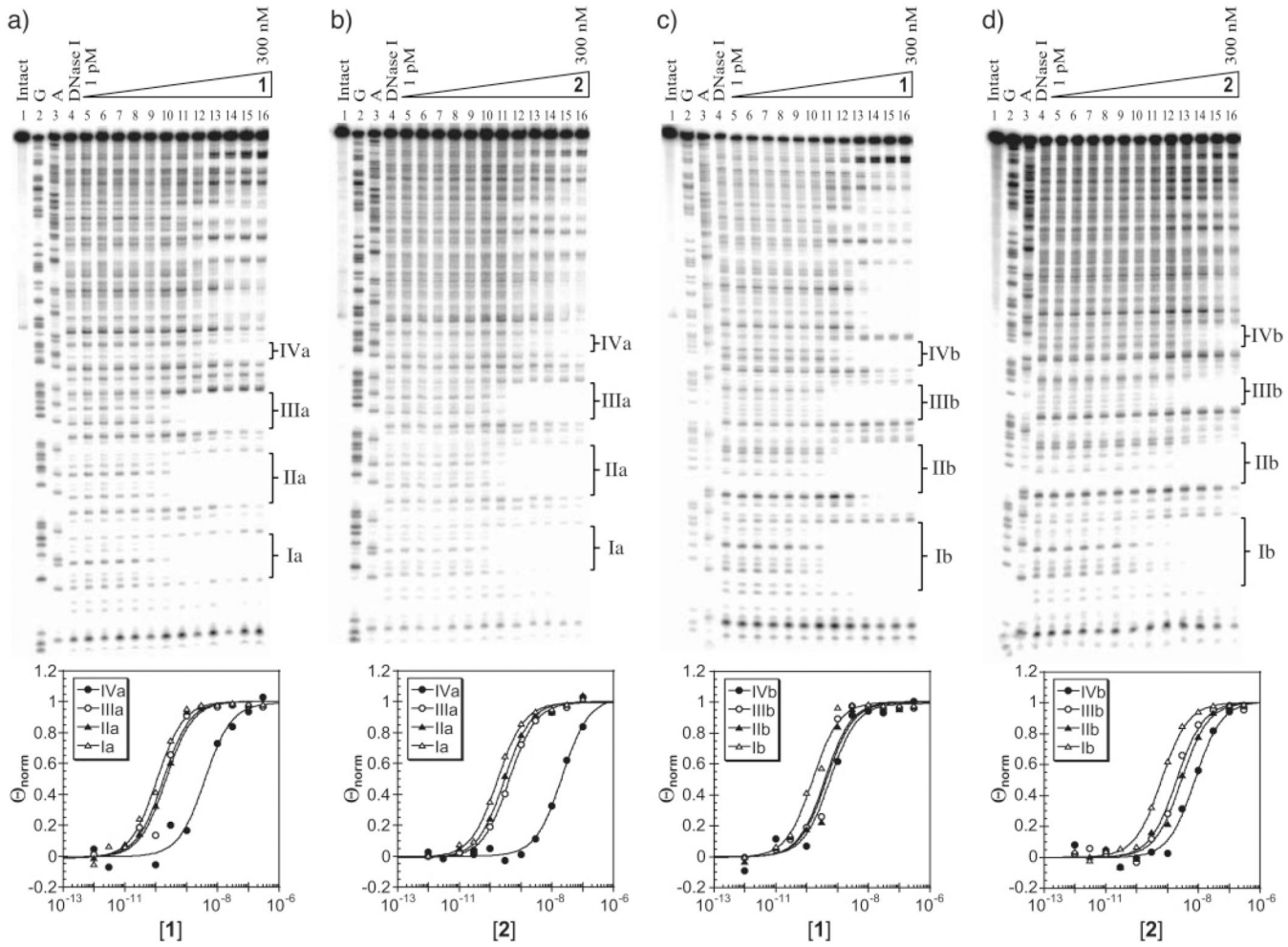


Figure 4.

DNase I footprinting gels and corresponding isotherms of polyamides **1** and **2** on pKAM3 and pKAM4. (a) Polyamide **1** on pKAM3. (b) Polyamide **2** on pKAM3. (c) Polyamide **1** on pKAM4. (d) Polyamide **2** on pKAM4.

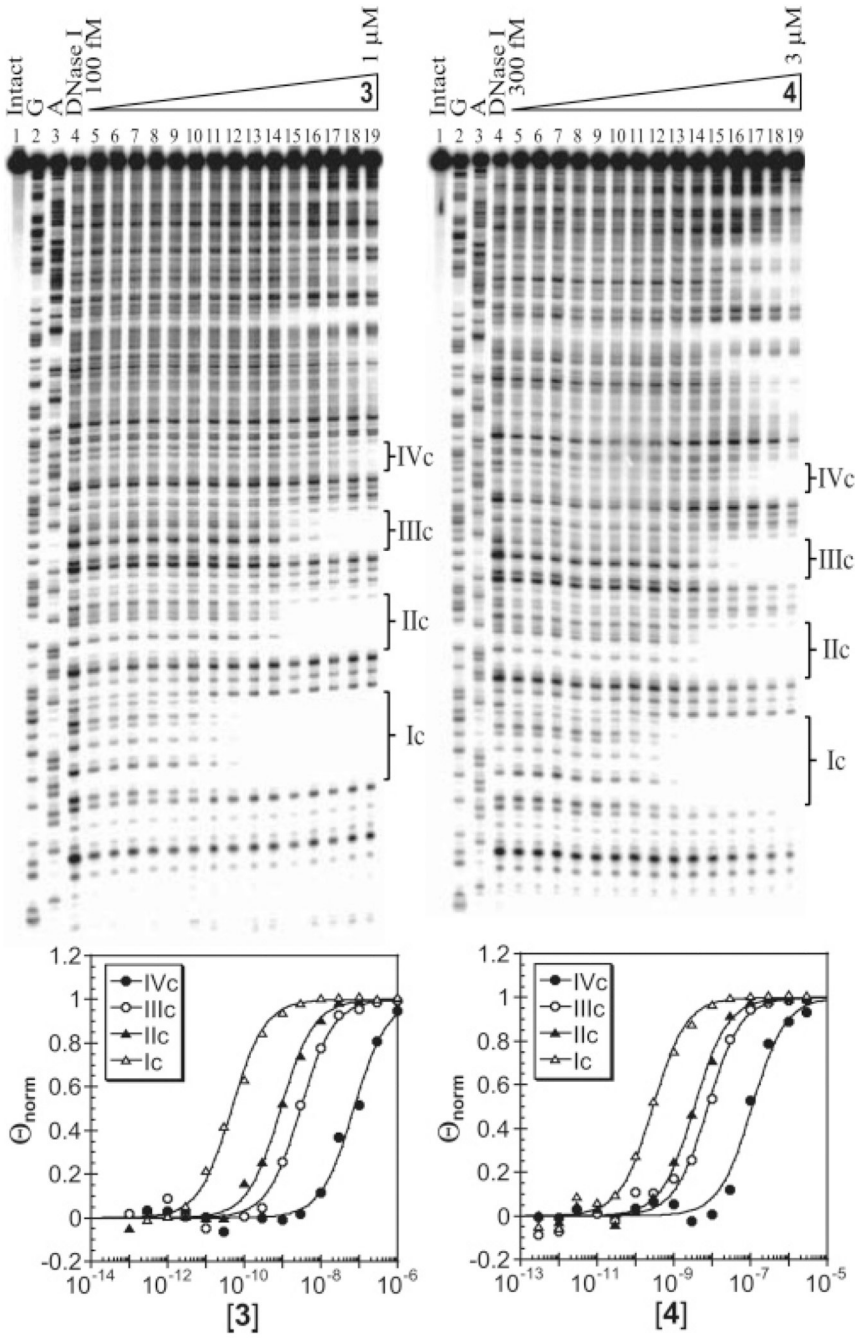


Figure 5.
DNase I footprinting gels and corresponding isotherms of polyamides **3** and **4** on pJWP17.

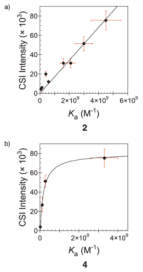


Figure 6.
CSI array intensities correlate well with DNase I footprinting-determined K_a values. (a) Polyamide **2** vs CSI array fit to eq 2. (b) Polyamide **4** vs CSI array fit to eq 1.

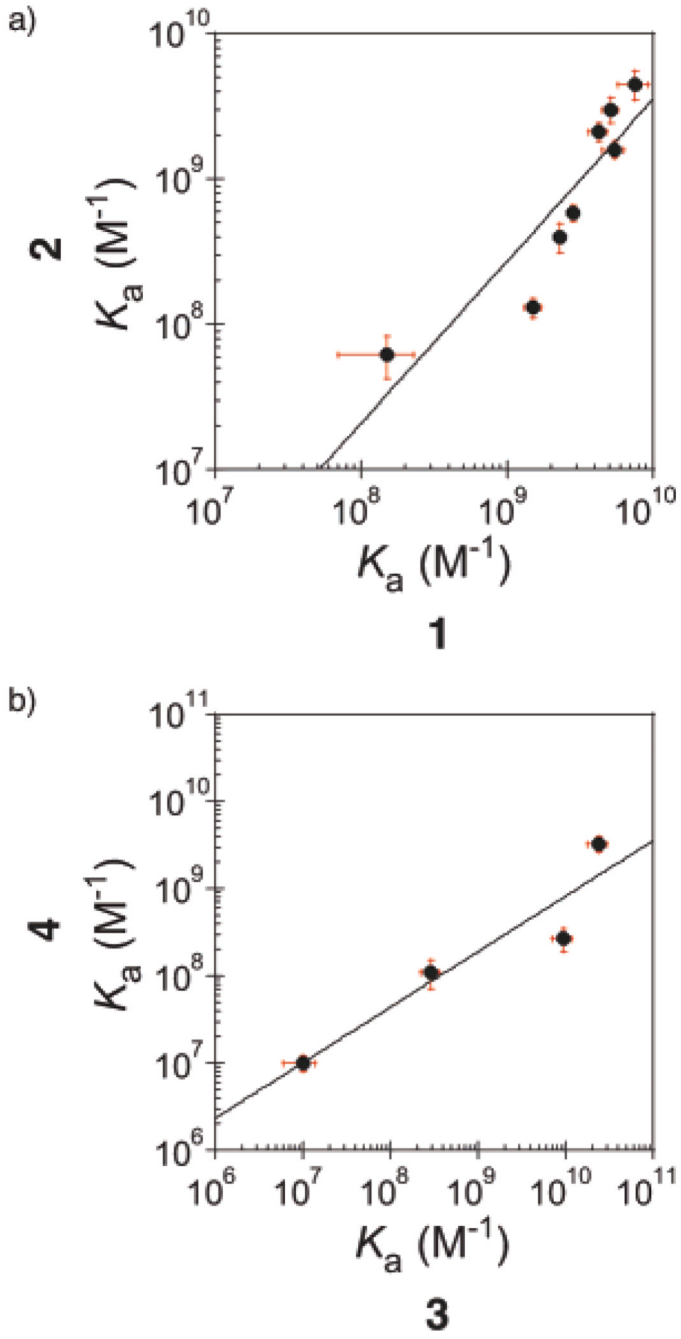


Figure 7.

(a) Correlation of K_a values for polyamide **1** (fluorescein labeled) and polyamide **2** (Cy3 labeled). (b) Correlation of K_a values for polyamide **3** (unlabeled) and polyamide **4** (Cy3 labeled).

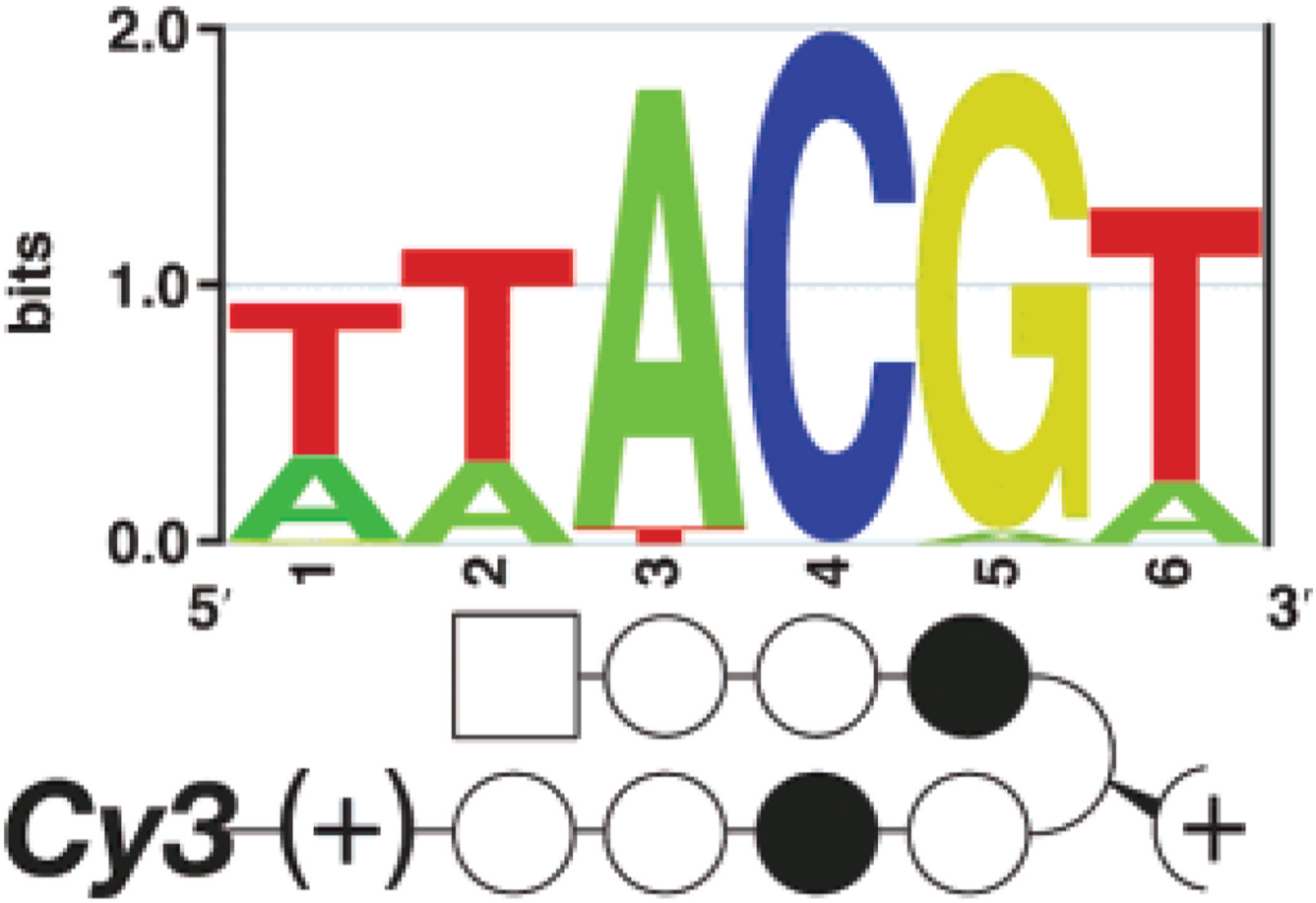


Figure 8.
Sequence logo for polyamide **2**.

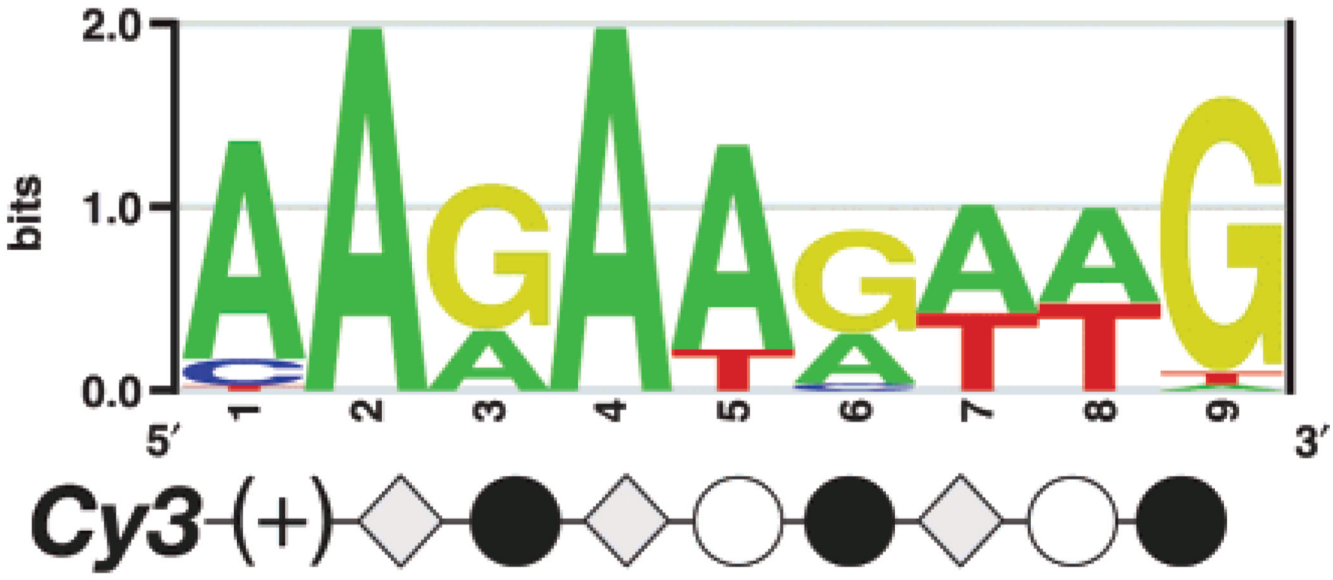
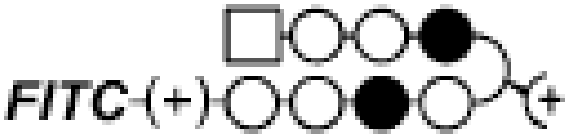
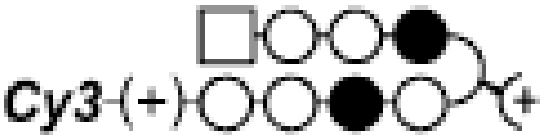


Figure 9.
Sequence logo for polyamide **4**.

Table 1

Quantitative DNase I Footprinting Derived K_a Values (M^{-1}) for Polyamides **1** and **2**, Their 10 Base Pair Binding Sites, and the Corresponding CSI Microarray Intensity[a]

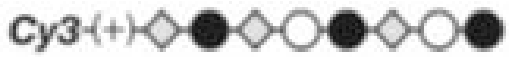
a) pKAM-3		**Ia**	**IIa**	**IIIa**	**IVa**
	Polyamide	**TTTTACGTAA**	**TTTTACGTAG**	**TTTTACGTGA**	**TTTTACGGAA**
1		$7.5 (\pm1.8) \times 10^9$ **[1]**	$5.1 (\pm0.6) \times 10^9$ **[1.5]**	$4.2 (\pm0.6) \times 10^9$ **[1.8]**	$1.5 (\pm0.8) \times 10^8$ **[50]**
2		$4.5 (\pm1.0) \times 10^9$ **[1]**	$3.0 (\pm0.6) \times 10^9$ **[1.5]**	$2.1 (\pm0.3) \times 10^9$ **[2.1]**	$6.2 (\pm2.0) \times 10^7$ **[73]**
	CSI Intensity (× 10^3)	75.6 (±9.9)	51.4 (±7.4)	31.3 (±4.8)	4.2 (±1.4)

b) pKAM-4		**Ib**	**IIb**	**IIIb**	**IVb**
	Polyamide	**TTTTACGTGA**	**AATTTCGTGT**	**GCTTTCGTCC**	**ACCTTCGTGA**
1		$5.4 (\pm0.9) \times 10^9$ **[1.3]**	$2.3 (\pm0.1) \times 10^9$ **[3.2]**	$2.8 (\pm0.2) \times 10^9$ **[2.6]**	$1.5 (\pm0.2) \times 10^9$ **[5]**
2		$1.6 (\pm0.2) \times 10^9$ **[2.8]**	$4.0 (\pm0.9) \times 10^8$ **[11]**	$5.8 (\pm0.7) \times 10^8$ **[7.8]**	$1.3 (\pm0.2) \times 10^8$ **[34]**
	CSI Intensity (× 10^3)	31.3 (±4.8)	20.0 (±2.8)	12.0 (±1.5)	6.0 (±0.4)

[a] All footprinting incubations were conducted at a minimum in triplicate at 23 °C for 14 h. Standard deviations are shown in parentheses. The bracketed numbers are $K_{a-max}/K_{a-current}$ to compare K_a values within each polyamide series.

Table 2

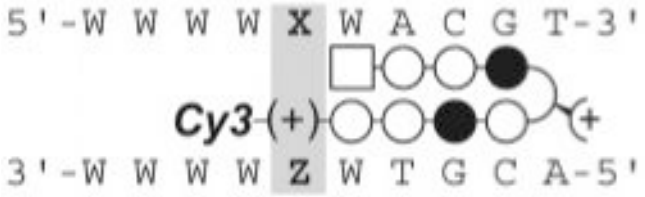
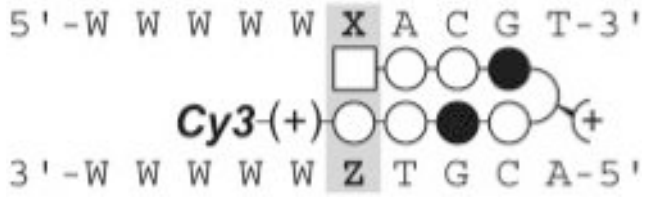
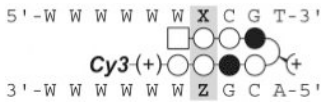
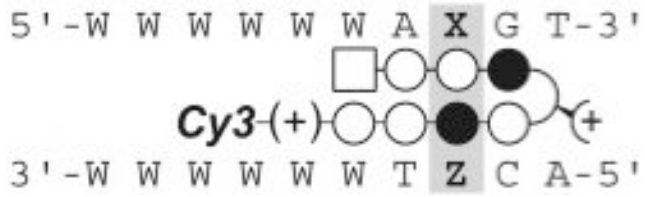
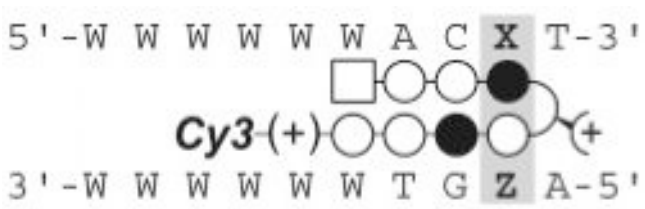
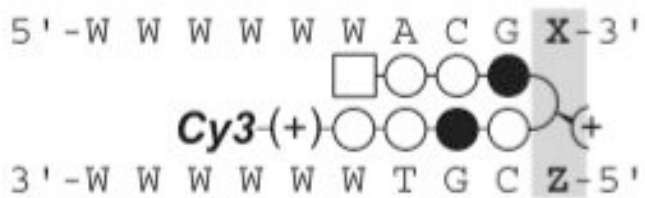
Quantitative DNase I Footprinting Derived K_a Values (M^{-1}) for Polyamides **3** and **4**, Their 10 Base Pair Binding Sites, and the Corresponding CSI Microarray Intensity[a]

pJWP-17		**Ic**	**IIc**	**IIIc**	**IVc**
	Polyamide	**AAGAAGAAGT**	**AAGAAGTTCA**	**ATGTTTGTTGA**	**ATGAAGACGA**
3		2.4 (±0.6) × 10^{10} **[1]**	9.3 (±2.3) × 10^{9} **[3]**	2.9 (±0.7) × 10^{8} **[80]**	1.0 (±0.4) × 10^{7} **[2400]**
4		3.3 (±0.7) × 10^{9} **[1]**	2.7 (±0.8) × 10^{8} **[10]**	1.1 (±0.4) × 10^{8} **[30]**	1.0 (±0.2) × 10^{7} **[330]**
	CSI Intensity (× 10^3)	75.2 (±9.2)	51.2 (±6.2)	26.8 (±7.3)	4.1 (±0.4)

[a] All footprinting incubations were conducted at a minimum in triplicate at 23 °C for 14 h. Standard deviations are shown in parentheses. The bracketed numbers are $K_{a-max}/K_{a-current}$ to compare K_a values within each row.

Table 3

Microarray-Derived Binding Affinities and Specificities of All Single Base Pair Mismatch Sites for Polyamide **2**[a]

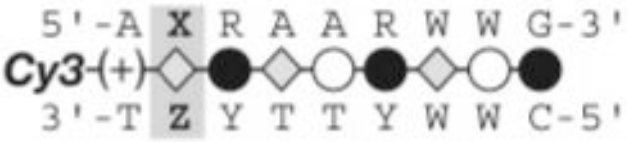
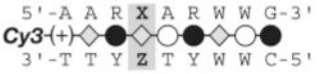
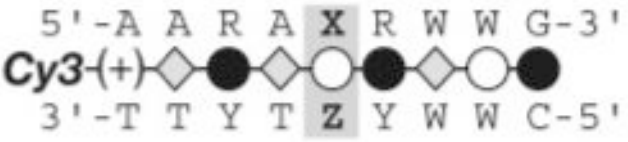
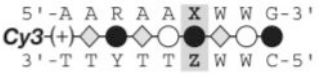
Polyamide 2	X·Z	K_a (M^{-1})
	A·T	2.0 (1.4) × 10^9
	T·A	2.5 (1.2) × 10^9
	C·G	6.9 (1.4) × 10^8
	G·C	6.8 (1.6) × 10^8
	A·T	1.8 (1.3) × 10^9
	T·A	2.7 (1.2) × 10^9
	C·G	1.0 (2.0) × 10^8
	G·C	1.3 (2.2) × 10^8
	A·T	2.2 (1.3) × 10^9
	T·A	1.1 (1.6) × 10^9
	C·G*	≤ 10^8
	G·C	1.3 (2.5) × 10^8
	A·T*	≤ 10^8
	T·A*	≤ 10^8
	C·G	2.2 (1.3) × 10^9
	G·C*	≤ 10^8
	A·T	1.2 (1.4) × 10^9
	T·A	2.9 (1.8) × 10^8
	C·G	2.4 (1.8) × 10^8
	G·C	2.2 (1.3) × 10^9
	A·T	1.3 (1.4) × 10^9
	T·A	2.2 (1.3) × 10^9
	C·G*	≤ 10^8
	G·C*	≤ 10^8

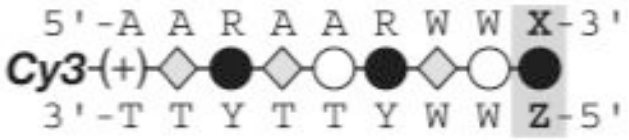
[a] All K_a values are derived from the geometric average of all CSI binding site intensities on the array containing a specified sequence, converted to a K_a value using eq 4, corrected to include an error term ε.[13] The values in parentheses are the geometric standard deviations for each K_a value.

X·Z entries marked with a superscripted "*" contain averaged intensities below ε. For these entries, an upper bound on the K_a is estimated based on the log−log plot of K_a versus intensity found in Supporting Information Figure 3.

Table 4

Microarray-Derived Binding Affinities and Specificities of All Single Base Pair Mismatch Sites for Polyamide **4**[a]

Polyamide 4	X·Z	K_a (M^{-1})
	A·T	1.6 (2.2) × 10^8
	T·A	8.7 (2.0) × 10^7
	C·G	4.3 (1.8) × 10^7
	G·C	4.7 (1.8) × 10^7
	A·T	1.6 (2.2) × 10^8
	T·A	7.7 (2.2) × 10^7
	C·G	2.1 (1.8) × 10^7
	G·C	3.2 (1.9) × 10^7
	A·T	1.3 (2.0) × 10^8
	T·A	3.5 (1.7) × 10^7
	C·G	8.4 (1.6) × 10^7
	G·C	2.0 (2.2) × 10^8
	A·T	1.6 (2.2) × 10^8
	T·A	4.3 (1.8) × 10^7
	C·G	9.9 (1.9) × 10^6
	G·C	7.5 (2.6) × 10^6
	A·T	1.6 (2.2) × 10^8
	T·A	8.3 (2.2) × 10^7
	C·G	9.1 (2.1) × 10^6
	G·C	1.1 (2.0) × 10^7
	A·T	1.5 (2.1) × 10^8
	T·A	5.5 (1.7) × 10^7
	C·G	8.5 (1.7) × 10^7
	G·C	1.7 (2.3) × 10^8
	A·T	1.7 (2.2) × 10^8
	T·A	1.5 (2.2) × 10^8
	C·G	2.2 (2.2) × 10^7
	G·C	2.0 (2.4) × 10^7
	A·T	1.6 (2.5) × 10^8
	T·A	1.5 (2.0) × 10^8
	C·G	3.5 (2.0) × 10^7
	G·C	3.9 (2.0) × 10^7

Polyamide 4	X·Z	K_a (M^{-1})
	A·T	1.2 (1.9) × 10^8
	T·A	1.1 (2.0) × 10^8
	C·G	1.1 (1.7) × 10^8
	G·C	1.6 (2.2) × 10^8

[a] All K_a values are derived from the geometric average of all CSI binding site intensities on the array containing a specified sequence, converted to a K_a value using eq 3, corrected to include an error term ε.[13] The values in parentheses are the geometric standard deviations for each K_a value.